Crucial Importance of Water-Entropy Effect for Predicting Hot Spots in Protein-Protein Complexes

Hiraku Oshima,[a] Satoshi Yasuda,[b] Takashi Yoshidome,[a] Mitsunori Ikeguchi,[c] and Masahiro Kinoshita*[a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

"Hot spots" are residues accounting for the majority of the protein-protein binding free energy (BFE) despite that they comprise only a small fraction of the protein-protein interface. A hot spot can be found experimentally by measuring the BFE change upon mutating it to alanine: The mutation gives rise to a significantly large increase in the BFE. Theoretical prediction of hot spots is an enthusiastic subject in biophysics, biochemistry, and bioinformatics. For the development of a reliable prediction method, it is essential to understand the physical origin of hot spots. To this end, we calculate the water-entropy gains upon the binding both for a wild-type complex and for its mutant complex using a hybrid method of the angle-dependent integral equation theory applied to a molecular model for water and the morphometric approach. We note that this type of calculation has never been employed in the previously reported methods. The BFE change due to alanine mutation is evaluated only from the change in the water-entropy gain with no parameters fitted to the experimental data. It is shown that the overall performance of predicting hot spots in our method is higher than that in Robetta, a standard free-energy-based method using fitting parameters, when the most widely used criterion for defining an actual hot spot is adopted. This result strongly suggests that the water-entropy effect we calculate is the key factor governing basic physics of hot spots.

1 Introduction

Protein-protein interactions are essential in a variety of biological processes within living cells and organisms. Thermodynamics of the interactions can be probed experimentally by alanine scanning mutagenesis [1–3]. Proteins interact with each other through an interface which consists of several interface residues. In alanine scanning mutagenesis, an interface residue is systematically replaced by alanine and the induced change in the binding free-energy (BFE), $\Delta\Delta G$, is experimentally measured. $\Delta\Delta G$ is defined as $\Delta G^{\rm mut} - \Delta G^{\rm wt}$, where $\Delta G^{\rm wt}$ and $\Delta G^{\rm mut}$ are the BFEs upon complex formation of the wild-type and alanine-mutated proteins, respectively. As alanine does not have a side chain beyond the β-carbon, the importance of each side-chain group in the binding can be estimated. According to a large number of experimental studies, the mutation of a small subset of interface residues leads to a significantly large increase in the BFE. These residues are called "hot spots" [1–3]. How to determine which residues are hot spots is a long-standing issue whose resolution would have significant implications for practical applications such as rational drug design and protein engineering [4,5]. Alanine scanning mutagenesis is an effective means of clarifying protein-protein interactions, but systematic identification of hot spots requires a large amount of experimental effort. In contrast, theoretical prediction of hot spots using computers is faster and its cost performance is higher. The theoretical prediction has thus become one of the most challenging subjects in biophysics, biochemistry, and bioinformatics. [6–18] Understanding the physical origin of hot spots is essential in the development of a reliable prediction method.

Our recent theoretical analyses based on a statistical-mechanical theory for fluids have shown that the water entropy is the key quantity in elucidating the folding/unfolding mechanisms of proteins [19–28]. Upon protein folding, for example, a large gain in the water entropy occurs for the following reason. As illustrated in Fig.1, the presence of a side chain generates an excluded space which the centers of water molecules cannot enter. The volume of the excluded space is referred to as "excluded volume" (EV). When side chains are closely packed or contact one another, the excluded spaces overlap and the total EV decreases by the volume of this overlapped space. This decrease provides an increase in the total volume available to the translational displacement (i.e., an increase in the number of possible coordinates of centers) of water molecules. This accompanies an increase in the number of accessible config-

[a] *Institute of Advanced Energy, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. E-mail: kinoshit@iae.kyoto-u.ac.jp*
[b] *Graduate School of Energy Science, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.*
[c] *Graduate School of Nanobioscience, Yokohama City University, 1-7-29, Suehirocho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan*

urations of water, leading to a gain of the water entropy. The importance of this water-entropy effect in protein folding and unfolding was argued in our earlier publications[29,30]. (In a strict sense, the water-entropy change upon protein folding is affected not only by the EV but also by three more geometric measures as described in Sec. 2.2.) We believe that this type of argument can be applied to the protein-protein binding. When side chains of residues are closely packed or contact one another in the protein-protein interface, the resultant overlap of the excluded volumes leads to a water-entropy gain. In particular, hot spots should make a large contribution to the gain. This concept is consistent with the experimentally known fact that the side chain of a hot spot is closely packed with side chains of the surrounding residues[5,31].

The water-entropy roles in the receptor-ligand binding have been discussed by significantly many authors in literature[32–36]. In most of the studies, they consider isobaric condition and look at only the contributions from the water near receptor and ligand surfaces. By contrast, we employ isochoric condition and argue the water-entropy effect by incorporating the water within a considerably larger length scale which is taken into account as the EV-dependent term. Let us take a solute insertion process as a simpler example. The hydration free energy is the same under isobaric and isochoric conditions, whereas the hydration entropy and energy are not[37]. Under isochoric condition, the water-entropy effect arising from the translational displacement of water molecules is suitably reflected in the hydration entropy the value of which is always negative. Under isobaric condition, by contrast, if the solute is hydrophobic and sufficiently large, for instance, the bulk water expands by the volume that is larger than the excluded volume of the solute[37]. As a consequence, the hydration entropy becomes positive. Likewise, upon the binding of hydrophobic receptor and ligand under isobaric condition, the compression of bulk water occurs, leading to water-entropy loss[38]. Of course, the water entropy always increases upon the binding under isochoric condition. We consider isochoric condition that is free from the effects of compression or expansion of bulk water on hydration thermodynamic quantities and under which the physical interpretation of the water-entropy change is straightforward.

Up to now, a variety of computational or theoretical methods have been proposed for predicting hot spots or protein-protein interactions: the machine learning method, method focused on the solvent-accessible surface area (ASA) of interface residues, method based on protein evolution, free-energy-based method with fitting parameters, and molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) method. (1) In the machine learning method[6–9], the features of interface residues such as atomic contacts, hydrogen bonds, and shapes of residues and the information about true (i.e., experimentally known) hot spots are used as the input training data for a learning machine model. In the training stage, a number of parameters entering the model are optimized so that the true hot spots can be predicted with the highest performance. On the basis of the matters thus learned, the model predicts new hot spots which are not treated in the training stage. (2) In the method focused on the ASA, the ASA changes for interface residues upon the protein-protein binding are used as principal parameters for the hot spot determination[10–12]. It is experimentally known that the solvent-accessible surface of a hot spot residue tends to be largely buried upon the binding. Tuncbag *et al.*[10] looked at the difference between hot spots and non-hot spots in terms of the distribution of the ASA change. They determined a threshold value of the ASA change: A residue is predicted to be a hot spot if its ASA change is larger than the threshold value and it is predicted to be a non-hot spot otherwise. (3) In the method based on protein evolution, protein-protein interactions are predicted by using the knowledge of protein evolution[39,40]. The useful knowledge is the following: Interface residues are found to be more conserved than the rest of surface residues, hot-spot residues are more conserved than the other interface residues[10,39,40]; and hot-spot residues are found to correlate with structurally conserved residues (i.e., residues conserved within the family with respect to the structural coordinates, disregarding sequence, and motif information[41]). It is also known that the prediction performance is substantially improved by serving this knowledge as an input feature for a machine learning method[42]. (4) In the free-energy-based method[13–15], a free-energy function based on the energy terms (i.e., van der Waals and electrostatic interactions and hydrogen bonds) and on the solvation free energy is employed for calculating $\Delta\Delta G$. These terms are combined linearly with weighting parameters determined for the best fit to the experimental data. Robetta[13] is a well established free-energy-based method which has become the *de facto* standard of comparison in the field, and it can freely be used as a web service called Robetta Server. In Robetta, the solvent effect is taken into account by the effective energy function 1 (EEF1) model[43], an implicit solvent model. (5) In the MM/PBSA method[16–18], representative conformations of the protein complex are constructed from snapshots along an explicit solvent molecular dynamics (MD) simulation. The free energy function comprises the molecular mechanics potential energy, solvation free energy, and conformational entropy. The solvation free energy is decomposed into electrostatic and nonpolar terms. The former term is estimated using the Poisson-Boltzmann equation treating the solvent as dielectric continuum and the latter one is assumed to be proportional to the ASA of the protein regardless of its structure[44]. No parameters fitted to the experimental data are employed.

The controlling parameters in methods (1), (2), and (4), which are determined from the experimental data, possess no

apparent physical meaning though they are useful for improving the prediction performance to a remarkable extent. In methods (1) through (3), the important solvent effects are not explicitly incorporated. We note that the water-entropy effect discussed in the second paragraph can be characterized by the following: It is reasonably taken into account only by a molecular model for water[45,46]; not only the water near the protein surface but also the water within a considerably larger length scale makes a substantial contribution to the entropic effect (i.e., the effect cannot be considered in terms of the ASA alone); the protein-water-water triplet and higher-order correlations play critical roles; and the solvation entropy of a protein is largely dependent on the details of its structure. Only our theoretical method wherein these factors are fully incorporated is capable of elucidating the great water-entropy gain upon apoplastocyanin folding experimentally estimated[19] and the microscopic mechanisms of pressure[24,25,47] and cold[27,28] denaturating of proteins. It is now obvious that methods (4) and (5) are not well qualified as theoretical tools fully accounting for the water-entropy effect.

The BFE consists of the entropic and energetic components. The entropic component comprises a water-entropy gain and a conformational-entropy loss of the proteins. The energetic component can be discussed primarily in terms of an energy decrease arising from protein-protein interactions and an energy increase due to the loss of protein-water hydrogen bonding and van der Waals interaction. We believe that the water-entropy gain predominates over the other factors in the components as a principal quantity governing basic physics of hot spots. Therefore, in the present study, we estimate the BFE only through the water-entropy gain upon the binding. The water-entropy gain is calculated under the isochoric condition which can be handled more readily, because the BFE is the same irrespective of the condition (isochoric and isobaric)[38,48]. The calculation is performed both for a wild-type complex and for its mutant complex using a hybrid method of the angle-dependent integral equation theory combined with the multipolar water model and the morphometric approach. The BFE change upon alanine mutation thus obtained is compared with the experimental data, and the discrimination between hot spots and non-hot spots is carried out. Despite that no fitting parameters are used in our method, the correlation coefficient determined from the comparison with the experimental data is almost the same as that in Robetta. Further, the overall performance of predicting hot spots in our method is higher than that in Robetta when the most widely used criterion for defining an actual hot spot is adopted. We also examine how often each amino-acid residue is predicted to be a hot spot by our method. As a result, the overall tendency that tryptophan, arginine, and tyrosine residues are frequently identified as hot spots[1], which is experimentally known, is well reproduced. These results suggest that the water-entropy effect we refer to is crucially important in the prediction of hot spots.

2 Model and theory

2.1 Water and protein models

A water molecule is modeled as a hard sphere with diameter $d_S = 0.28$ nm in which a point dipole and a point quadrupole of tetrahedral symmetry are embedded[49,50]. The influence of molecular polarizability of water is included by employing the self-consistent mean field (SCMF) theory[49,50]. At the SCMF level the many-body induced interactions are reduced to pairwise additive potentials involving an effective dipole moment. The effective dipole moment thus determined is about 1.42 times larger than the bare gas-phase dipole moment. The absolute temperature T is set at 298 K. The number density of the bulk water ρ_S at this temperature is taken to be that of real water on the saturation curve, $\rho_S d_S^3 = 0.7317$.

We calculate the hydration entropy (HE), S, which represents the loss of the water-entropy when a protein with a prescribed structure is immersed in water. (When it is emphasized that the solvent is water, "solvation" is replaced by "hydration".) The change in the water-entropy from structure A to structure B is obtained as the HE of structure B minus that of structure A. We model a protein as a set of fused hard spheres. This is reasonable because the HE is not significantly dependent on the protein-water interaction potentials. Imai *et al.* considered the native structures of a total of eight peptides and proteins and calculated S using three-dimensional reference interaction site model (3D-RISM) theory combined with the all-atom potentials and the SPC/E water model[51]. Even when the protein-water electrostatic potentials, which are quite strong, are shut off and only the Lennard-Jones (LJ) potentials are retained, $|S|$ decreases merely by less than 5%. Modeling a protein as a set of fused hard spheres can also be justified as follows. The hydration free energy μ, entropy S, and energy U under the isochoric condition are calculated for a spherical solute with diameter 0.28 nm using the angle-dependent integral equation theory[37,48,52–62] combined with the multipolar water model. For the hard-sphere solute with zero charge, the calculated values are μ=5.95 k_BT, S=−9.22 k_B, and U=−3.27 k_BT. Here, k_B is Boltzmann's constant. When the point charge $-0.5e$ (e is the electronic charge) is embedded at its center, the calculated values are μ=−32.32 k_BT, S=−10.11 k_B, and U=−42.43 k_BT. Thus, S is fairly insensitive to the solute-water interaction potential while μ and U are largely influenced by it. This insensitivity of the hydration entropy to the solute-water interaction comes from the fact that the contribution of water molecules near the surface of the solute is sufficiently small.

We consider protein-protein complexes whose structures

have been solved by X-ray crystallography and for which the data of the BFE change upon alanine mutation are available from the Alanine Scanning Energetics Database (ASEdb)[1,2] and the dataset used in Ref. 6. The structures are obtained from the Protein Data Bank (PDB). We assume that the protein structures in a complex and in a monomer are the same and that no structural changes are induced upon the mutation[1]. The two complexes used in Ref. 6, 1fc2 and 1jtg, are excluded because the structure in the unbound state significantly differs from that in the bound state[63–65]. We calculate the BFE change upon the mutation only for the residues in the protein-protein interface (i.e., interface residues). An interface residue is defined as a residue with $\Delta\text{ASA} \geq 1$ Å^2[8] where ASA is the solvent-accessible surface area calculated with a probe sphere of radius 1.4 Å and ΔASA represents the ASA decrease due to burial of the residue upon the protein-protein binding. A residue in the vicinity of the interface with $\text{ASA} = 0$ Å^2 before and after the binding are also defined as an interface residue. The mutation of glycine or proline is excluded because it could change conformational flexibility of the protein backbone, causing a significant difference between the wild-type and mutant structures. We eventually consider 341 mutants of 18 complexes in the present study (Table.1).

To remove unrealistic overlaps of the protein atoms, the structure of the wild-type complex is modified by the energy minimization using the CHARMM biomolecular simulation program[66] through the Multi-scale Modeling Tools in Structural Biology (MMTSB) program[67]. The structure of the mutant complex is obtained by replacing the atoms beyond β-carbon in the side chain of a residue by hydrogen atoms. Each complex is then divided into two isolated proteins, partners 1 and 2. We thus obtain the six structures of the wild-type complex, its partners 1 and 2, mutant complex, and its partners 1 and 2. The (x,y,z) coordinates of all the protein atoms in the backbone and side chains are used as part of the input data to account for the characteristics of each structure on the atomic level. The diameter of each atom is set at the σ-value of the Lennard-Jones potential parameters of CHARMM22[66].

2.2 Morphometric approach to hydration entropy of a protein

Since a molecular model is employed for water, the angle-dependent version[48,55,56,59,62], which is explained in Sec.2.3, must be used for the integral equation theory, an elaborate statistical-mechanical theory. However, its extension to complex solute molecules such as proteins is rather difficult due to the mathematical complexity. This problem can be overcome by combing it with the morphometric approach[68,69]. The idea of the approach is to express a hydration quantity such as S by the linear combination of four geometric measures of a solute molecule,

$$S/k_{\rm B} = C_1 V_{\rm ex} + C_2 A + C_3 X + C_4 Y. \quad (1)$$

Here, $V_{\rm ex}$ is the EV, A is the ASA, and X and Y are the integrated mean and Gaussian curvatures of the solvent-accessible surface, respectively. We calculate these measures by means of an extension[68] of Connolly's algorithm[70,71]. Contributions to X and Y from the lines of intersecting spheres and those to Y from the points where three lines meet are also included in the calculation[68].

The idea of the morphometric form expressed by Eq.(1) is that details of the solute shape enters $S/k_{\rm B}$ via the four geometric measures. Therefore, the four coefficients in the linear combination can be determined in simple geometries. They are determined from calculations of the hydration thermodynamic quantity for spherical solutes with various diameters. The hydration thermodynamic quantities for the spherical solutes are calculated using the angle-dependent integral equation theory described in Sec.2.3. The morphometric form applied to the spherical solutes reduces to

$$S/k_{\rm B} = C_1\left(\frac{4\pi}{3}d_{\rm US}^3\right) + C_2\left(4\pi d_{\rm US}^2\right) + 4\pi C_3 d_{\rm US} + 4\pi C_4, \quad (2)$$

where $d_{\rm US} = (d_{\rm U} + d_{\rm S})/2$ and $d_{\rm U}$ is the solute diameter. (Hereafter, the subscripts "S" and "U" represent "water (solvent)" and "solute", respectively.) The four coefficients are determined using the least square fitting to Eq.(2). Once they are determined, the hydration thermodynamic quantity of a protein with any structure is obtained by calculating only its four geometric measures.

The high reliability of the morphometric approach has already been demonstrated in our earlier publications[19,25,29,45,46,68]. For example, the results from the three-dimensional integral equation theory[72,73] applied to the same model protein immersed in a simple solvent (the solvent particles interact through strongly attractive potential such as water molecules) can be reproduced with sufficient accuracy by the morphometric approach where the four coefficients are determined in the manner explained above. By a hybrid of the angle-dependent integral equation theory combined with the multipolar water model and the morphometric approach, the experimentally measured changes in thermodynamic quantities upon apoplastocyanin folding are quantitatively reproduced[19]. Moreover, great progresses have been made in elucidating the microscopic mechanisms of pressure[24,25,47], cold[27,28], and heat[26] denaturating of proteins, in discriminating the native fold from a number of misfolded decoys[29,74–76], and in uncovering the rotational mechanism of F_1-ATPase[77] by our theoretical methods in which the morphometric approach is combined with the integral equation theory or its angle-dependent version. We note that even for a large protein

complex with a fixed structure the calculation of the geometric measures and the HE is finished in less than ∼10 seconds on a workstation.

We evaluate the BFE change upon alanine mutation of a hot spot or non-hot spot by accounting for the water-entropy effect alone. The difference, HE of a complex − (HE of partner 1 + HE of partner 2), represents the water-entropy gain upon the protein-protein binding. The gains for the wild-type (wt) and the mutant (mut) are expressed as follows:

$$\Delta S^{\mathrm{wt}} = S^{\mathrm{wt}}_{\mathrm{complex}} - S^{\mathrm{wt}}_{\mathrm{partner1}} - S^{\mathrm{wt}}_{\mathrm{partner2}}, \tag{3a}$$

$$\Delta S^{\mathrm{mut}} = S^{\mathrm{mut}}_{\mathrm{complex}} - S^{\mathrm{mut}}_{\mathrm{partner1}} - S^{\mathrm{mut}}_{\mathrm{partner2}}. \tag{3b}$$

The change in the water-entropy gain upon the binding caused by the mutation is given as

$$\Delta\Delta S = \Delta S^{\mathrm{mut}} - \Delta S^{\mathrm{wt}}. \tag{4}$$

In many cases the mutation decreases the degree of packing in the protein-protein interface, inducing a decrease in the water-entropy gain upon the binding and a negative value of $\Delta\Delta S$. In particular, its absolute becomes quite large if a closely packed residue (i.e., a hot spot) is mutated. We compare the change in the water-entropy gain calculated by our method, $-\Delta\Delta S_{\mathrm{calc}}/k_{\mathrm{B}} = \Delta\Delta G_{\mathrm{calc}}/(k_{\mathrm{B}}T)$, with the experimentally observed BFE change, $\Delta\Delta G_{\mathrm{obs}}$.

2.3 Angle-dependent integral equation theory

A spherical solute is considered in the analysis using the angle-dependent integral equation theory[37,48,52–62]. A feature of this theory is that the water-water and solute-water orientational correlations are explicitly taken into account. A protein can be treated when this theory is combined with the morphometric approach described in Sec.2.2. Hard spheres of diameter d_{U} (i.e., solutes) are immersed in the model water. The solute-water interaction potential is expressed as

$$u_{\mathrm{US}}(r) = \infty \qquad \text{for} \quad r < d_{\mathrm{US}}, \tag{5a}$$

$$u_{\mathrm{US}}(r) = 0 \qquad \text{for} \quad r \geq d_{\mathrm{US}}. \tag{5b}$$

The Ornstein-Zernike (OZ) equation for the mixture comprising water molecules and spherical solutes can be written as

$$\eta_{\alpha\beta}(12) = \frac{1}{8\pi^2}\sum_{\gamma}\rho_{\gamma}\int c_{\alpha\gamma}(13)\{\eta_{\gamma\beta}(32) + c_{\gamma\beta}(32)\}d3, \tag{6a}$$

$$\eta_{\alpha\beta}(12) = h_{\alpha\beta}(12) - c_{\alpha\beta}(12); \qquad \alpha,\beta = \mathrm{S},\mathrm{U}, \tag{6b}$$

where h and c are the total and direct correlation functions, respectively, (ij) represents $(\boldsymbol{r}_{ij}, \Omega_i, \Omega_j)$, $\boldsymbol{r}_{ij}$ is the vector connecting the centers of particles i and j, Ω_i denotes the three Euler angles describing the orientation of particle i, $\int d3$ represents integration over all position and angular coordinates of particle 3, and ρ is the number density. The closure equation is expressed by[78]

$$c_{\alpha\beta}(12) = \int_{r_{12}}^{\infty}\Big[h_{\alpha\beta}(12)\frac{\partial\{w_{\alpha\beta}(12) - b_{\alpha\beta}(12)\}}{\partial r'_{12}}\Big]dr'_{12} - u_{\alpha\beta}(12)/(k_{\mathrm{B}}T) + b_{\alpha\beta}(12), \tag{7a}$$

$$w_{\alpha\beta}(12) = -\eta_{\alpha\beta}(12) + u_{\alpha\beta}(12)/(k_{\mathrm{B}}T), \tag{7b}$$

where u is the pair potential, b the bridge function, T the absolute temperature, and $r_{12} = |\boldsymbol{r}_{12}|$. In the present analysis, the hypernetted-chain (HNC) approximation is employed ($b = 0$). As proved in earlier work[48], the hydration free energy can be calculated with quantitative reliability using the theory combined with the multipolar model for water, despite the neglect of the bridge function. We assume that the solutes are immersed in water at infinite dilution ($\rho_{\mathrm{U}} = 0$). The calculation process can then be split into two steps:

Step (i). Solve Eqs.(6) and (7) for bulk water. Calculate the correlation functions X_{SS} ($X = h, c$).

Step (ii). Solve Eqs.(6) and (7) for the solute-water system using the correlation functions obtained in step (i) as input data. Calculate the correlation functions X_{US} ($X = h, c$).

For the numerical solution of Eqs.(6) and (7), the pair potentials and correlation functions are expanded in a basis set of rotational invariants, and the basic equations are reformulated in terms of the projections $X^{mnl}_{\mu\nu}(r)$ occurring in the rotational-invariant expansion of X[37,48,52–62]. The expansion considered for m, $n \leq n_{\max} = 4$ gives sufficiently accurate results for hard-sphere solutes. The basic equations are then numerically solved using the robust, highly efficient algorithm developed by Kinoshita and coworkers[55–57]. In the numerical treatment, a sufficiently long range r_{L} is divided into N grid points ($r_i = i\delta r$, i=0, 1, ..., $N-1$; $\delta r = r_{\mathrm{L}}/N$) and all of the projections are represented by their values on these points. The grid width and the number of grid points are set at $\delta r = 0.01 d_{\mathrm{S}}$ and $N = 4096$, respectively.

The hydration free energy μ is calculated using the Morita-Hiroike formula extended to molecular liquids[19,48,79,80]:

$$\mu/(k_{\mathrm{B}}T) = \frac{\rho_{\mathrm{S}}}{(8\pi^2)}\iiiint 4\pi\Big[\frac{1}{2}\{h_{\mathrm{US}}(r,\theta,\phi,\chi)\}^2 - \frac{1}{2}h_{\mathrm{US}}(r,\theta,\phi,\chi)c_{\mathrm{US}}(r,\theta,\phi,\chi) - c_{\mathrm{US}}(r,\theta,\phi,\chi)\Big]r^2\sin\theta dr d\theta d\phi d\chi, \tag{8}$$

where $h_{\mathrm{US}} = g_{\mathrm{US}} - 1$ and the integration range is [0, ∞] for r, [0, π] for θ, and [0, 2π] for ϕ and χ. The hydration entropy

S is evaluated through the numerical differentiation of μ with respect to the temperature [19,48] as

$$S = -\left(\frac{\partial \mu}{\partial T}\right)_{\mathrm{V}} = -\frac{\mu(T+\delta T)-\mu(T-\delta T)}{2\delta T}, \ \ \delta T = 5\mathrm{K}. \tag{9}$$

The high reliability of the angle-dependent integral equation theory combined with the multipolar water model has been demonstrated in a variety of studies. As proved in our earlier work[48], the angle-dependent integral equation theory is far superior to the RISM and related theories [78,81,82] in analyses on the hydrophobic hydration. The former gives a quantitatively accurate value of the hydration free energy of a nonpolar solute. The dielectric constant of bulk water calculated using the angle-dependent integral equation theory combined with the multipolar water model is ∼83 that is in good agreement with the experimental value ∼78. The hydration properties of hydrophilic solutes can well be reproduced [83]. The theory has been quite successful in elucidating the water structure near uncharged, charged, and metal surfaces [48–50,83].

2.4 Performance evaluation

We define a residue as an actual hot spot when $\Delta\Delta G_{\mathrm{obs}} \geq \theta_{\mathrm{obs}}$ holds. We examine the two values: $\theta_{\mathrm{obs}} = 2$ kcal/mol (criterion 1) and $\theta_{\mathrm{obs}} = 1$ kcal/mol (criterion 2). The former is a standard criterion employed in several studies previously reported, and the latter has been adopted in the original implementation of Robetta [13]. It is not advisable to employ only a single value for θ_{obs} because the typical errors of the experimental data are as large as ± 0.5 kcal/mol [1]. On the theoretical side, a residue is predicted to be a hot spot when $\Delta\Delta G_{\mathrm{calc}} \geq \theta_{\mathrm{calc}}$ holds. As a result, there are four possible outcomes: In the case where a residue is an actual hot spot, it is counted as a true positive (TP) when it is predicted to be a hot spot and as a false negative (FN) when it is predicted as a non-hot spot; in the case where a residue is an actual non-hot spot, it is counted as a true negative (TN) when it is predicted to be a non-hot spot and as a false positive (FP) when it is predicted to be a hot spot. The performance measures for a prediction method are then defined as follows:

$$\begin{aligned} \mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ \mathrm{Specificity} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \text{F-measure} &= \frac{2\times\mathrm{Recall}\times\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}}. \end{aligned} \tag{10}$$

"Recall", which is defined for the actual hot spots, represents the proportion of those which are successfully predicted to be hot spots while "precision" is defined for the predicted hot spots and is the proportion of those which are actually hot spots. "Specificity", which is defined for the actual non-hot spots, represents the proportion of those which are successfully predicted to be non-hot spots. "F-measure" is the harmonic average of "recall" and "precision". These measures are in the range from 0 to 1, and a larger value implies better performance.

We look at the Receiver Operating Characteristic (ROC) curve [84] as another measure of the prediction performance. An ROC curve is a graph representing a trade-off of "recall" and "1 − specificity". Here, "recall" and "1 − specificity" represent proportions of the true prediction for hot spots and of the false prediction for non-hot spots, respectively. For a given value of θ_{calc}, a single point ("1 − specificity", "recall") on the curve is determined. The ROC curve is traced by varying θ_{calc} from ∞ to $-\infty$. On the point (0,0) all the residues are predicted to be non-hot spots due to $\theta_{\mathrm{calc}} = \infty$. On the point (1,1) all the residues are predicted to be hot spots due to $\theta_{\mathrm{calc}} = -\infty$. If the ROC curve of a prediction method passes through points with higher "recall" and lower "1 − specificity" (i.e., if the curve is closer to (0,0) → (0,1) → (1,1)), it is generally regarded as a better predictor because the probability of successful prediction for hot spots is better irrespective of the value of θ_{calc} set. The area under the ROC curve (AUC) can be used as a quality metric. "AUC = 1" represents a perfect predictor while "AUC = 0.5" displays the performance that is as bad as the random chance. The AUC is equivalent to the probability that a prediction method ranks a randomly chosen hot spot higher than a randomly chosen non-hot spot in respect of the BFE change calculated [84].

We also determine the best value of θ_{calc} as the value minimizing the balanced error rate (BER) defined as

$$\mathrm{BER} = \frac{1}{2}\left(\frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}\right), \tag{11}$$

where the BER represents the averaged error rate of a prediction method. The performance measures ("recall", "precision", "specificity", and "F-measure") are calculated at the best value of θ_{calc}.

3 Results and discussion

3.1 Correlation between calculated and experimental results

We compare the results from our method with those from Robetta [13] for the same set of mutants. Robetta is a well established free-energy-based method which has become the *de facto* standard of comparison in the field, and it can freely be used as a web service called Robetta Server. In Robetta,

global optimization of the hydrogen bonding network is performed just for a wild-type complex (no other optimization for the backbone and side chains is performed[85]). We upload the protein structures which are not optimized by our method (i.e., the raw data obtained from the PDB) to Robetta Server and obtain the results from Robetta ($\Delta\Delta G_{\mathrm{Robetta}}$) in its original way.

The observed BFE change, $\Delta\Delta G_{\mathrm{obs}}$, is plotted against the value calculated by our method in Fig.2(a) or that by Robetta in Fig.2(b). Since we consider only the water-entropy effect, a quantitative comparison between our calculated value with $\Delta\Delta G_{\mathrm{obs}}$ is not available, and the calculated value is presented as $-\Delta\Delta S_{\mathrm{calc}}/k_{\mathrm{B}}$. In Robetta, the scale of the vertical line quantitatively corresponds to that of the horizontal line, which makes the quantitative comparison available. This is not surprising because Robetta employs the parameter fitting to the experimental results. Our major concern is the correlation coefficient R calculated from Fig.2: $R = 0.522$ in our method and $R = 0.515$ in Robetta. Even though in our method only the water entropy effect is taken into account with no parameter fitting, R from our method is as high as that from Robetta.

3.2 Performance of prediction method

We compare the performances of our method and Robetta using the ROC analysis. The usefulness of the ROC analysis is explained in Sec.2.4. The ROC curves with criteria 1 and 2 are shown in Fig.3(a) and in Fig.3(b), respectively. With criterion 1, "recall" on the curve of our method is higher than that of Robetta for given "$1-$ specificity" except in the region of very high θ_{calc}. With criterion 2, the two curves share almost the same characteristics. The AUC-values of the ROC curves are given in Table 2. With criterion 1 the AUC from our method is higher than that from Robetta. With criterion 2 the AUC-values from the two methods are almost the same.

We determine the best value of θ_{calc} for our method and Robetta. The performance measures, recall, precision, specificity, and F-measure, at the best value are listed in Tables 3 and 4. With criterion 1 (Table.3), the indices of our method are all higher than those of Robetta, and our method exhibits higher performance of the hot-spot prediction. With criterion 2 (Table 4), recall and F-measure are larger in our method while precision and specificity are larger in Robetta. Overall, the two methods share almost the same performance. We note that criterion 1 is a standard criterion adopted in several previous studies (i.e., the most widely used criterion). In summary, our method exhibits higher overall performance than Robetta when the standard criterion is employed. Even with the particular criterion (criterion 2) chosen in the original implementation of Robetta, the two methods share almost the same overall performance.

The free-energy function of Robetta is expressed as a linear combination of energetic terms and solvation free-energy terms[13]. The solvation free-energy terms are estimated on the basis of an implicit (dielectric continuum) model for water. The coefficients in the linear combination are controlled so that the free-energy function provides the best fit to the experimental data. Such parameter fitting usually leads to remarkable improvement of the method. In our method, by contrast, only the water-entropy effect is taken into account and no fitting parameters are employed. It is surprising that nevertheless the overall performance provided by our method is higher than that of Robetta. These results strongly suggest that this water-entropy effect, which can reasonably be taken into account only by employing a molecular model for water, plays the key role in the protein-protein binding and presents crucial importance in the hot-spot prediction.

3.3 Packing efficiencies around a hot-spot residue before and after alanine mutation

We consider the two hot-spot residues, W304 of 1a22 and K52 of 1jrh. For these residues, both $\Delta\Delta G_{\mathrm{obs}}$ and $-\Delta\Delta S_{\mathrm{calc}}/k_{\mathrm{B}}$ are quite large, and the water-entropy effect makes an essential contribution to the BFE change upon the mutation. Their points in the plot of $\Delta\Delta G_{\mathrm{obs}}$ against $-\Delta\Delta S_{\mathrm{calc}}/k_{\mathrm{B}}$ in Fig.2 (a) are almost on the linearly fitted line. 1a22 consists of human growth hormone (chain A) and human growth hormone binding protein (chain B), and the mutation for W304 in chain B induces the very large BFE change, 4.5 kcal/mol. We look at the interface residues around W304 before and after the mutation (Fig.4(a)). It is found that the interface residues around W304 are closely packed but the mutation of W304 by alanine gives rise to appreciable loss of this high packing efficiency. 1jrh consists of antibody A6 (chains L and H) and interferon γ receptor (chain I), and the mutation for K52 of chain I gives $\Delta\Delta G_{\mathrm{obs}} = 3.0$ kcal/mol. From Fig.4(b), the interface residues around K52 are also closely packed before the mutation while this is not true after the mutation. Since the loss of such close packing causes a large reduction in the total volume available to the translational displacement of water molecules, the mutation causes a large decrease in the water-entropy gain upon the binding. Thus, the generally known feature, a hot-spot residue is closely packed with the surrounding residues in the protein-protein interface, can be elucidated in terms of the water-entropy effect which is suitably incorporated in our method.

3.4 Amino-acid species frequently acting as hot spots

Using ASEdb, Bogan *et al.*[1] have shown that tryptophan, arginine, and tyrosine act as hot spots most frequently in the real systems. To check if our method can reproduce this

tendency, we investigate how often each amino-acid species is predicted to be a hot spot. We calculate the distribution, $N_i^{\rm hot}/\sum N_i^{\rm hot}$, where amino-acid species i is predicted as a hot spot $N_i^{\rm hot}$ times and the summation is taken for all the amino-acid species. The same calculation is also made for Robetta. In the hot-spot prediction, the best values of $\theta_{\rm calc}$ determined in Sec.3.2 (see Tables 3 and 4) are employed for our method and for Robetta, respectively, and the two different criteria, ($\theta_{\rm obs} = 2$ and 1 kcal/mol), are considered. The distributions calculated in our method and in Robetta are compared with the experimentally observed distribution in Fig.5 as well as in Tables 5 and 6. In the dataset we use, tyrosine (Y), tryptophan (W), lysine (K), and aspartic acid (D) act as hot spots most frequently. The frequencies of arginine (R) and glutamic acid (E) are also significantly high. In both of our method and Robetta, tyrosine (Y), tryptophan (W), arginine (R), and lysine (K) exhibit especially high frequencies. Fairly high frequencies are counted for glutamic acid (E) and aspartic acid (D) in our method and for glutamic acid (E) and glutamine (Q) in Robetta. These results are independent of the hot-spot criterion (i.e., the value of $\theta_{\rm obs}$). Our method as well as Robetta has turned out to be capable of reproducing the overall tendency of the experimental results. It is interesting to note that the distributions in our method and in Robetta are quite similar. Again, it is indicated that the water-entropy effect plays the key role in the protein-protein binding.

4 Factors to be taken into account in future work

4.1 Dehydration-penalty effect

First, we define "dehydration penalty"[29] upon the protein-protein binding which is a primary energetic constituent. The binding leads to an energy decrease arising from the gain of protein-protein interactions but accompanies an energy increase caused by the loss of protein-water interactions. Even when electrostatic attractive interactions between a group in the protein with a positive charge and water oxygens and between a group with a negative charge and water hydrogens are lost upon the binding, the resulting energy increase is compensated if the two groups are in close contact with each other in the protein-protein interface. However, such a contact cannot always be formed, leading to "dehydration penalty" that is an increase in the system energy. Further, the binding usually accompanies the loss of conformational entropies of the proteins. The BFE comprises the water-entropy gain, dehydration penalty, and conformational-entropy loss. The BFE change upon the mutation can be written as

$$\Delta\Delta G_{\rm calc} = -T\Delta\Delta S_{\rm calc} + \Delta\Delta\Lambda - T\Delta\Delta S_{\rm conf},$$

where $\Delta\Delta\Lambda = \Delta\Lambda^{\rm mut} - \Delta\Lambda^{\rm wt}$ and $\Delta\Delta S_{\rm conf} = \Delta S_{\rm conf}^{\rm mut} - \Delta S_{\rm conf}^{\rm wt}$ denote changes in the dehydration penalty and in the conformational-entropy loss, respectively.

To discuss the dehydration-penalty effect first, let us consider two example cases. (The conformational-entropy effect is discussed in a later subsection.) In the first case, residue A with a negatively charged side chain, which is in contact with residue B with a positively charged side chain in the wild-type complex, is mutated by alanine. Residues A and B belong to partners 1 and 2, respectively. The wild-type complex undergoes no dehydration penalty upon the binding thanks to the compensation described above, and $\Delta\Lambda^{\rm wt}$ is almost zero. For the complex formed after the mutation, however, the loss of electrostatic attractive interactions between the negatively charged side chain of residue A and water hydrogens is not necessarily compensated. If the compensation is not attained, $\Delta\Lambda^{\rm mut}$ takes a large, positive value with the result of $\Delta\Delta\Lambda > 0$. In the second case, there are no positively charged side chains in the immediate vicinity of the negatively charged side chain of residue A in the wild-type complex. It is obvious that $\Delta\Lambda^{\rm wt}$ is positive and large. The dehydration penalty vanishes if residue A is mutated by alanine, leading to $\Delta\Delta\Lambda < 0$. We intend to incorporate the energetic component into our method in terms of the dehydration-penalty effect in future work. In the first case the incorporation leads to upward shift of a point plotted in Fig.2(a) while in the second case it leads to downward shift.

4.2 Structural changes caused by protein-protein binding and alanine mutation

We consider the result for Y94 of 1dan whose plot in Fig.2(a) deviates largely from the fitted line. 1dan is a complex consisting of factor VIIA (chains L and H) and tissue factor (chains T and U). As shown in Fig.6, the structure of factor VIIA in the wild-type complex is significantly elongated. Factor VIIA isolated is not likely to take such an elongated structure (i.e., it should take a compact structure) due to the water-entropy effect. The elongated structure should be stabilized only after the complex formation with its partner. Nevertheless, we assume that the structure of factor VIIA remains unchanged even when they are separated from each other. This assumption can give rise to a significant error of the BFE change evaluated as $-\Delta\Delta S_{\rm calc}/k_{\rm B}$ when a residue of factor VIIA in the protein-protein interface is mutated by alanine. In fact, the correlation coefficient of our method improves to 0.543 if all the mutants of 1dan (64 mutants) are removed from the plot in Fig.2(a).

In general, it is possible that the structures of the complex and one of partners 1 and 2 undergo significant changes upon alanine mutation. In such cases, the BFE change calculated includes a significant error. If the structural changes caused by the protein-protein binding and/or alanine mutation can be es-

timated, the performance of a prediction method for hot spots is expected to improve to a large extent. In the previous subsection, we commented on the potential importance of the energetic component, namely, protein-protein and protein-water electrostatic interactions neglected in our method. They are included in the free-energy function of Robetta. Nevertheless, as observed from the comparison between Fig.2(a) and Fig.2(b), the points of Robetta and our method plotted for charged residues (a good example is D51A of 1bxi) exhibit similar deviations from the fitted lines. The similarity is suggestive that the method of incorporating the effects of protein-protein and protein-water electrostatic interactions in Robetta is to be revisited and that the structural changes caused by the protein-protein binding and/or alanine mutation, which are not taken into account in both of the two methods, can be more substantial than the energetic component.

4.3 Conformational-entropy effect

The change in the conformational entropy upon the protein-protein binding is defined as

$$\Delta S_{\rm conf} = S_{\rm conf,complex} - S_{\rm conf,partner1} - S_{\rm conf,partner2}.$$

$\Delta S_{\rm conf}$ is usually negative due to the reduction in the degree of freedom for structural fluctuation of side chains of the interface residues. The reduction should be remarkable for side chains of hot spots because they become closely packed with side chains of the surrounding residues upon the binding. Thus, the mutation of a hot spot by alanine leads to a significantly large, positive value of $\Delta\Delta S_{\rm conf} = \Delta S_{\rm conf}^{\rm mut} - \Delta S_{\rm conf}^{\rm wt}$. For a hot spot, our method gives a large, positive value of $-T\Delta\Delta S_{\rm calc}$. However, this value is reduced if the free-energy decrease arising from the conformational-entropy effect is added. Even for a non-hot spot, it is better to account for the effect because it could make a significantly large contribution to the BFE change upon the mutation.

5 Conclusions

We have investigated basic physics of hot spots in the interface of a protein-protein complex. Upon the protein-protein binding, the total volume available to the translational displacement of water molecules increases, leading to increases in the number of accessible configurations of water and a corresponding gain in the water entropy. It is shown that hot spots account for the majority of this water-entropy gain. Hence, the mutation of a hot spot by alanine results in a large increase in the binding free energy (BFE). This picture is consistent with the experimentally known feature of a hot spot[5,31] that its side chain is closely packed with side chains of the surrounding residues. We demonstrate that the water-entropy effect described above, which can reasonably be taken into account only by employing a molecular model for water, is crucially important in developing a hot-spot prediction method. Since the water-entropy effect has not been considered on the same level in any of the previously reported methods, the present study sheds new light on the hot-spot prediction.

We have tested a method in which the BFE upon the binding and the BFE change upon alanine mutation are obtained using the water-entropy term alone. The water-entropy gains upon the binding for wild-type and mutant complexes are calculated with the application of the angle-dependent integral equation theory for molecular liquids. The calculation of the water-entropy gain, which can be a formidable task when a molecular model is adopted for water, is accomplished with minor computational effort by combining the theory with the morphometric approach. The BFE change upon the mutation from our method is compared with the experimental data. No parameters fitted to the data enter our method. The correlation coefficient calculated from the comparison is almost the same as that of Robetta, a popular free-energy-based method using fitting parameters that has become the *de facto* standard of comparison in the field. The performance of our method is higher than that of Robetta when the most widely used criterion for defining an actual hot spot is adopted. These results manifest the validity of our physical picture treating the water-entropy effect as the key factor. If we further incorporate the additional factors such as the energetic component introduced in terms of the dehydration-penalty effect, structural changes caused by the protein-protein binding and alanine mutation, and conformational-entropy effect, the performance of our method will improve to a large extent. Works in this direction are in progress.

Acknowledgments

This work was supported by Grants-in-Aid for Scientific Research on Innovative Areas (Nos. 20118004 and 21118519), that on Priority Areas (No. 18074004), and that on (B) (Nos. 22300100 and 22300102) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, by the Grand Challenges in Next-Generation Integrated Simulation of Nanoscience and Living Matter, a part of the Development and Use of the Next-Generation Supercomputer Project of MEXT, by Kyoto University Global Center of Excellence (GCOE) of Energy Science, and by Kyoto University Pioneering Research Unit for Next Generation.

References

1 A. A. Bogan and K. S. Thorn, *J. Mol. Biol.*, 1998, **280**, 1.
2 K. S. Thorn and A. A. Bogan, *Bioinformatics*, 2001, **17**, 284.
3 T. Clackson and J. A. Wells, *Science*, 1995, **267**, 383.

4 W. L. DeLano, *Curr. Opin. Struct. Biol.*, 2002, **12**, 14.
5 I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Proteins Struct. Funct. Bioinf.*, 2007, **68**, 803.
6 S. Lise, C. Archambeau, M. Pontil and D. T. Jones, *BMC Bioinf.*, 2009, **10**, 365.
7 S. J. Darnell, D. Page and J. C. Mitchell, *Proteins Struct. Funct. Bioinf.*, 2007, **68**, 813.
8 K. Cho, D. Kim and D. Lee, *Nucleic Acids Res.*, 2009, **37**, 2672.
9 R. H. Higa and C. L. Tozzi, *Genet. Mol. Biol.*, 2009, **32**, 626.
10 N. Tuncbag, A. Gursoy and O. Keskin, *Bioinformatics*, 2009, **25**, 1513.
11 N. Tuncbag, O. Keskin and A. Gursoy, *Nucleic Acids Res.*, 2010, **38**, W402.
12 S. Grosdidier and J. Fernández-Recio, *BMC Bioinf.*, 2008, **9**, 447.
13 T. Kortemme and D. Baker, *Proc. Nat. Acad. Sci. U. S. A.*, 2002, **99**, 14116.
14 R. Guerois, J. E. Nielsen and L. Serrano, *J. Mol. Biol.*, 2002, **320**, 369.
15 A. Benedix, C. M. Becker, B. L. de Groot, A. Caflisch and R. A. Böckmann, *Nat. Methods*, 2009, **6**, 3.
16 I. Massova and P. A. Kollman, *J. Am. Chem. Soc*, 1999, **121**, 8133.
17 S. Huo, I. Massova and P. A. Kollman, *J. Comput. Chem.*, 2002, **23**, 15.
18 C. M. Reyes and P. A. Kollman, *J. Mol. Biol.*, 2000, **295**, 1.
19 T. Yoshidome, M. Kinoshita, S. Hirota, N. Baden and M. Terazima, *J. Chem. Phys.*, 2008, **128**, 225104.
20 Y. Harano and M. Kinoshita, *Chem. Phys. Lett.*, 2004, **399**, 342.
21 Y. Harano and M. Kinoshita, *Biophys. J.*, 2005, **89**, 2701.
22 Y. Harano and M. Kinoshita, *J. Phys. Condens. Matter*, 2006, **18**, L107.
23 Y. Harano and M. Kinoshita, *J. Chem. Phys.*, 2006, **125**, 024910.
24 Y. Harano, T. Yoshidome and M. Kinoshita, *J. Chem. Phys.*, 2008, **129**, 145103.
25 T. Yoshidome, Y. Harano and M. Kinoshita, *Phys. Rev. E*, 2009, **79**, 011912.
26 K. Amano, T. Yoshidome, Y. Harano, K. Oda and M. Kinoshita, *Chem. Phys. Lett.*, 2009, **474**, 190.
27 T. Yoshidome and M. Kinoshita, *Phys. Rev. E*, 2009, **79**, 030905(R).
28 H. Oshima, T. Yoshidome, K. Amano and M. Kinoshita, *J. Chem. Phys.*, 2009, **131**, 205102.
29 T. Yoshidome, K. Oda, Y. Harano, R. Roth, Y. Sugita, M. Ikeguchi and M. Kinoshita, *Proteins Struct. Funct. Bioinf.*, 2009, **77**, 950.
30 S. Yasuda, T. Yoshidome, H. Oshima, R. Kodama, Y. Harano and M. Kinoshita, *J. Chem. Phys.*, 2010, **132**, 065105.
31 O. Keskin, B. Ma and R. Nussinov, *J. Mol. Biol.*, 2005, **345**, 1281.
32 P. Cozzini, M. Fornabaio, A. Marabotti, D. J. Abraham, G. E. Kellogg and A. Mozzarelli, *Curr. Med. Chem.*, 2004, **11**, 3093.
33 M. Fuxreiter, M. Mezei, I. Simon and R. Osmany, *Biophys. J.*, 2005, **89**, 903.
34 R. Baron, P. Setny and J. A. McCammon, *J. Am. Chem. Soc.*, 2010, **132**, 12091.
35 P. Setny, R. Baron and J. A. McCammon, *J. Chem. Theory Comput.*, 2010, **6**, 2866.
36 G. Hummer, *Nat. Chem.*, 2010, **2**, 906.
37 N. M. Cann and G. N. Patey, *J. Chem. Phys.*, 1997, **106**, 8165.
38 M. Kinoshita, Y. Harano and R. Akiyama, *J. Chem. Phys.*, 2006, **125**, 244504.
39 A. C. F. Lewis, R. Saeed and C. Deane, *Mol. Biosyst.*, 2010, **6**, 55.
40 M. N. Wass, A. David and M. J. E. Sternberg, *Curr. Opin. Struct. Biol.*, 2011, **21**, 382.
41 B. Ma, T. Elkayam, H. Wolfson and R. Nussinov, *Proc. Nat. Acad. Sci. U. S. A.*, 2003, **100**, 5772.
42 Y. Ofran and B. Rost, *PLoS Comput. Biol.*, 2007, **3**, e119.
43 T. Lazaridis and M. Karplus, *Proteins Struct. Funct. Genet.*, 1999, **35**, 133.
44 D. Sitkoff, K. A. Sharp and B. Honig, *J. Phys. Chem.*, 1994, **98**, 1978.
45 M. Kinoshita, *Front. Biosci.*, 2009, **14**, 3419.
46 M. Kinoshita, *Int. J. Mol. Sci.*, 2009, **10**, 1064.
47 T. Yoshidome and M. Kinoshita, *Chem. Phys. Lett.*, 2009, **477**, 211.
48 M. Kinoshita, *J. Chem. Phys.*, 2008, **128**, 024507.
49 P. G. Kusalik and G. N. Patey, *J. Chem. Phys.*, 1988, **88**, 7715.
50 P. G. Kusalik and G. N. Patey, *Mol. Phys.*, 1988, **65**, 1105.
51 T. Imai, Y. Harano, M. Kinoshita, A. Kovalenko and F. Hirata, *J. Chem. Phys.*, 2006, **125**, 024911.
52 P. H. Fries and G. N. Patey, *J. Chem. Phys.*, 1985, **82**, 429.
53 G. M. Torrie, P. G. Kusalik and G. N. Patey, *J. Chem. Phys.*, 1988, **88**, 7826.
54 D. R. Bérard and G. N. Patey, *J. Chem. Phys.*, 1991, **95**, 5281.
55 M. Kinoshita and M. Harada, *Mol. Phys.*, 1993, **79**, 145.
56 M. Kinoshita and M. Harada, *Mol. Phys.*, 1994, **81**, 1473.
57 M. Kinoshita and D. R. Bérard, *J. Comput. Phys.*, 1996, **124**, 230.
58 M. Kinoshita, S. Iba and M. Harada, *J. Chem. Phys.*, 1996, **105**, 2487.
59 M. Kinoshita, *J. Solution Chem.*, 2004, **33**, 661.
60 M. Kinoshita, *J. Mol. Liq.*, 2005, **119**, 47.
61 M. Kinoshita, N. Matubayasi, Y. Harano and M. Nakahara, *J. Chem. Phys.*, 2006, **124**, 024512.
62 M. Kinoshita, *Condens. Matter Phys.*, 2007, **10**, 387.
63 J. Deisenhofer, *Biochemistry*, 1981, **20**, 2361.
64 H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata and I. Shimada, *Biochemistry*, 1992, **31**, 9665.
65 D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym and G. Schreiber, *Proc. Nat. Acad. Sci. U. S. A.*, 2005, **102**, 57.
66 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. A. Karplus, *J. Comput. Chem.*, 1983, **4**, 187.
67 M. Feig, J. Karanicolas and C. Brooks, *J. Mol. Graphics Modell.*, 2004, **22**, 377.
68 R. Roth, Y. Harano and M. Kinoshita, *Phys. Rev. Lett.*, 2006, **97**, 078101.
69 P. M. König, R. Roth and K. R. Mecke, *Phys. Rev. Lett.*, 2004, **93**, 160601.
70 M. L. Connolly, *J. Appl. Crystallogr.*, 1983, **16**, 548.
71 M. L. Connolly, *J. Am. Chem. Soc.*, 1985, **107**, 1118.
72 M. Ikeguchi and J. Doi, *J. Chem. Phys.*, 1995, **103**, 5011.
73 M. Kinoshita, *J. Chem. Phys.*, 2002, **116**, 3493.
74 Y. Harano, R. Roth and M. Kinoshita, *Chem. Phys. Lett.*, 2006, **432**, 275.
75 Y. Harano, R. Roth, Y. Sugita, M. Ikeguchi and M. Kinoshita, *Chem. Phys. Lett.*, 2007, **437**, 112.
76 S. Yasuda, T. Yoshidome, Y. Harano, R. Roth, H. Oshima, K. Oda, Y. Sugita, M. Ikeguchi and M. Kinoshita, *Proteins*, in press.
77 T. Yoshidome, Y. Ito, M. Ikeguchi and M. Kinoshita, *J. Am. Chem. Soc.*, 2011, **133**, 4030.
78 J. P. Hansen and I. R. Mcdonald, *Theory of Simple Liquids 3rd ed.*, Academic, London, 2006.
79 T. Morita, *Prog. Theor. Phys.*, 1960, **23**, 829.
80 T. Morita and K. Hiroike, *Prog. Theor. Phys.*, 1961, **25**, 537.
81 J. S. Perkyns and B. M. Pettitt, *Chem. Phys. Lett.*, 1992, **190**, 626.
82 J. S. Perkyns and B. M. Pettitt, *J. Chem. Phys.*, 1992, **97**, 7656.
83 M. Kinoshita and T. Yoshidome, *J. Chem. Phys.*, 2009, **130**, 144705.
84 T. Fawcett, *Pattern Recognit. Lett.*, 2006, **27**, 861.
85 T. Kortemme, D. Kim and D. Baker, *Sci. STKE*, 2004, **2004**, pl2.
86 *The PyMOL Molecular Graphics System, Version 1.2r2, Schrödinger, LLC*, 2009.

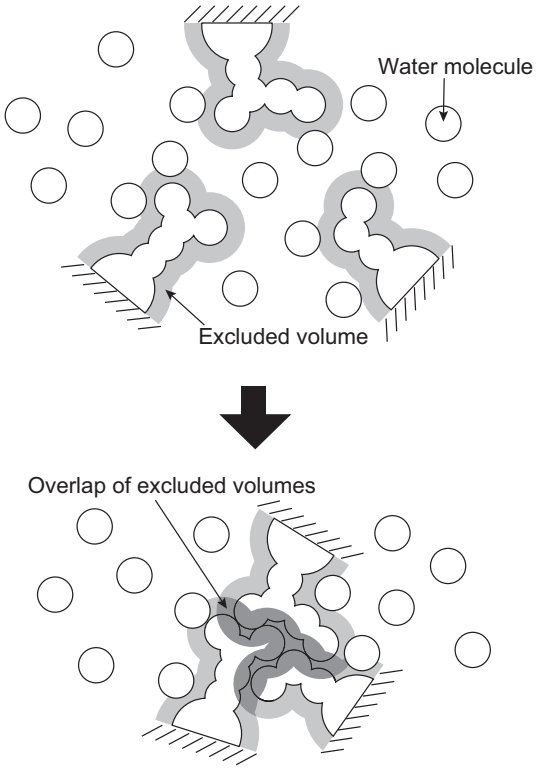

Fig. 1 Close packing of side chains of a protein. The overlap of excluded volumes leads to an increase in the total volume available to the translational displacement of water molecules.

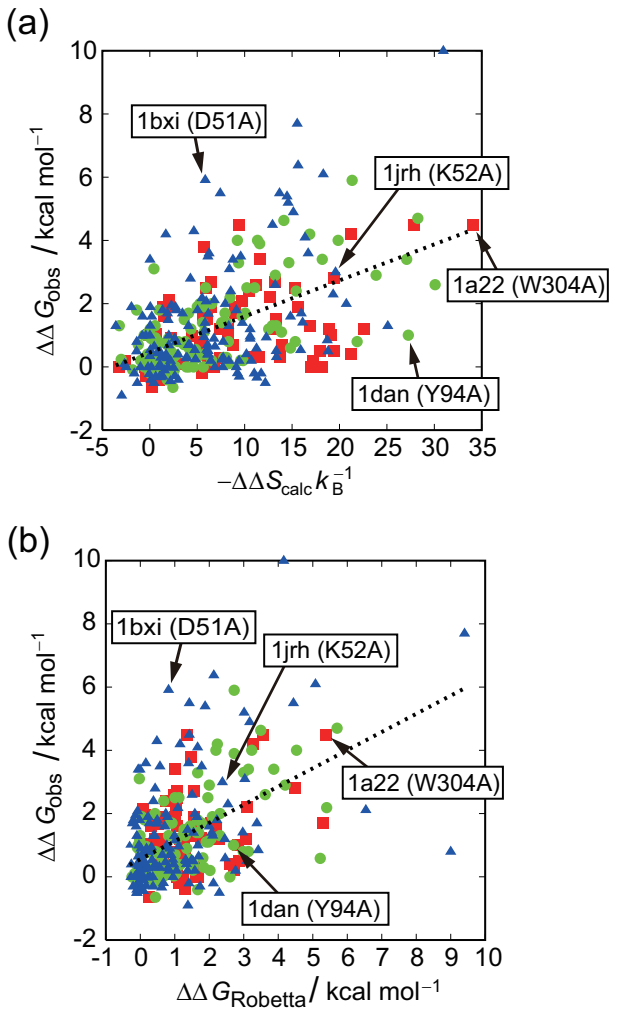

Fig. 2 Comparison between calculated and experimentally observed values of changes in the binding free energy upon alanine mutation. Red square, green circle, and blue triangle represent nonpolar, polar, and charged residues, respectively. The dashed lines represent linearly fitted lines. (a) Our result ($R = 0.522$). (b) Robetta's result ($R = 0.515$).

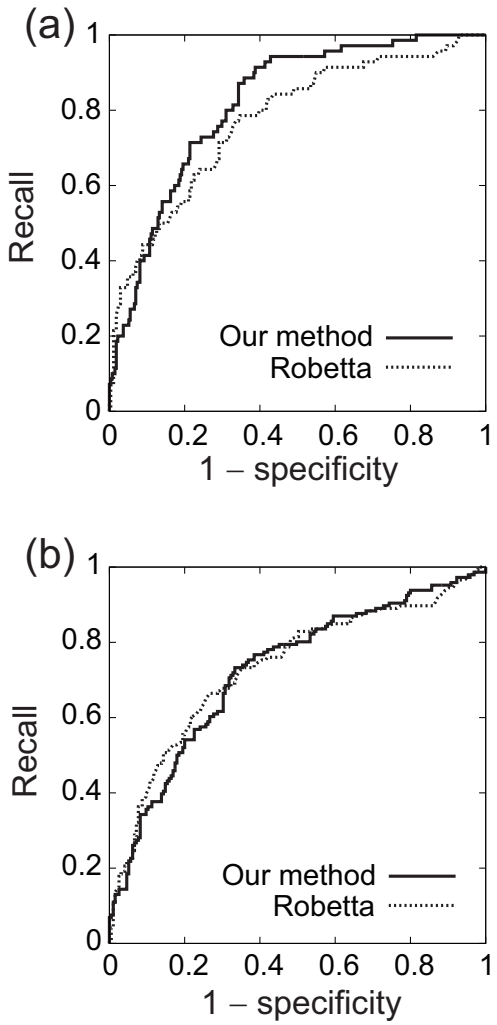

Fig. 3 ROC curves of our and Robetta's results. Solid lines: Our results. Dotted lines: Robetta's results. (a) Criterion 1 ($\theta_{\mathrm{obs}} = 2$ kcal/mol). (b) Criterion 2 ($\theta_{\mathrm{obs}} = 1$ kcal/mol).

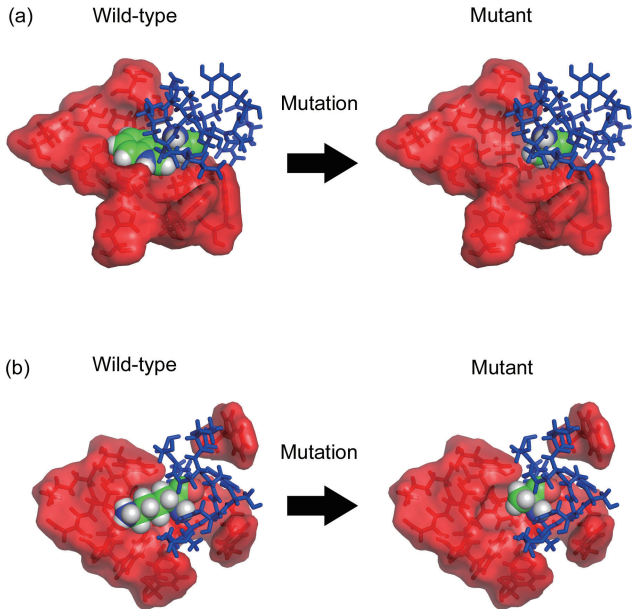

Fig. 4 Packing of interface residues around a hot spot in wild-type complex and that around alanine in mutant complex. We show only the atoms whose centers are within 6 Å from the center of the hot spot. (a) Chain A (in red) and chain B (in blue) of 1a22 are represented by molecular surface and sticks, respectively. W304 (a hot spot) in chain B is shown by a van der Waals representation. (Some residues of chain B are eliminated.) (b) Chain L (in red) and chain I (in blue) of 1jrh are represented by molecular surface and sticks, respectively. K52 (a hot spot) in chain I is shown by a van der Waals representation. The images are created by program PyMOL[86].

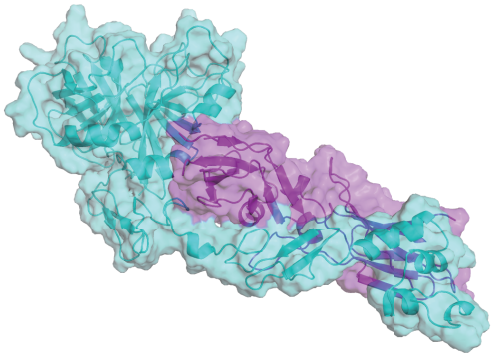

Fig. 6 Factor VIIA (in cyan) complexed with tissue factor (in magenta). Ribbon and molecular-surface representations are employed. The image is created by program PyMOL[86].

Table 2 Area under the Receiver Operating Characteristic (ROC) curve, AUC, calculated from criterion 1 ($\theta_{obs} = 2$ kcal/mol) or criterion 2 ($\theta_{obs} = 1$ kcal/mol)

	Criterion 1	Criterion 2
Our method	0.820	0.725
Robetta	0.774	0.731

Table 5 Distribution representing how frequently each amino-acid species acts as a hot spot in the case of criterion 1 ($\theta_{obs} = 2$ kcal/mol). The distribution is experimentally measured or theoretically predicted (our method or Robetta).

Amino-acid species	Experimental	Our method	Robetta
Val (V)	0.029	0.033	0.034
Leu (L)	0.014	0.033	0.034
Ile (I)	0.043	0.039	0.027
Met (M)	0	0.013	0.007
Phe (F)	0.014	0.046	0.040
Trp (W)	0.129	0.130	0.141
Ser (S)	0.014	0.006	0.013
Thr (T)	0.014	0.013	0.034
Asn (N)	0.057	0.039	0.047
Gln (Q)	0.029	0.052	0.081
Tyr (Y)	0.214	0.156	0.181
Cys (C)	0	0	0
Lys (K)	0.143	0.110	0.081
Arg (R)	0.100	0.143	0.121
His (H)	0.014	0.039	0.034
Asp (D)	0.129	0.071	0.060
Glu (E)	0.057	0.078	0.067

Table 6 Distribution representing how frequently each amino-acid species acts as a hot spot in the case of criterion 2 ($\theta_{obs} = 1$ kcal/mol). The distribution is experimentally measured or theoretically predicted (our method or Robetta).

Amino-acid species	Experimental	Our method	Robetta
Val (V)	0.027	0.035	0.034
Leu (L)	0.027	0.029	0.034
Ile (I)	0.048	0.047	0.027
Met (M)	0	0.012	0.007
Phe (F)	0.041	0.041	0.040
Trp (W)	0.123	0.13	0.141
Ser (S)	0.014	0.006	0.013
Thr (T)	0.034	0.017	0.034
Asn (N)	0.055	0.041	0.047
Gln (Q)	0.034	0.052	0.081
Tyr (Y)	0.185	0.163	0.181
Cys (C)	0	0	0
Lys (K)	0.096	0.105	0.081
Arg (R)	0.089	0.134	0.121
His (H)	0.027	0.035	0.034
Asp (D)	0.110	0.070	0.060
Glu (E)	0.089	0.081	0.067

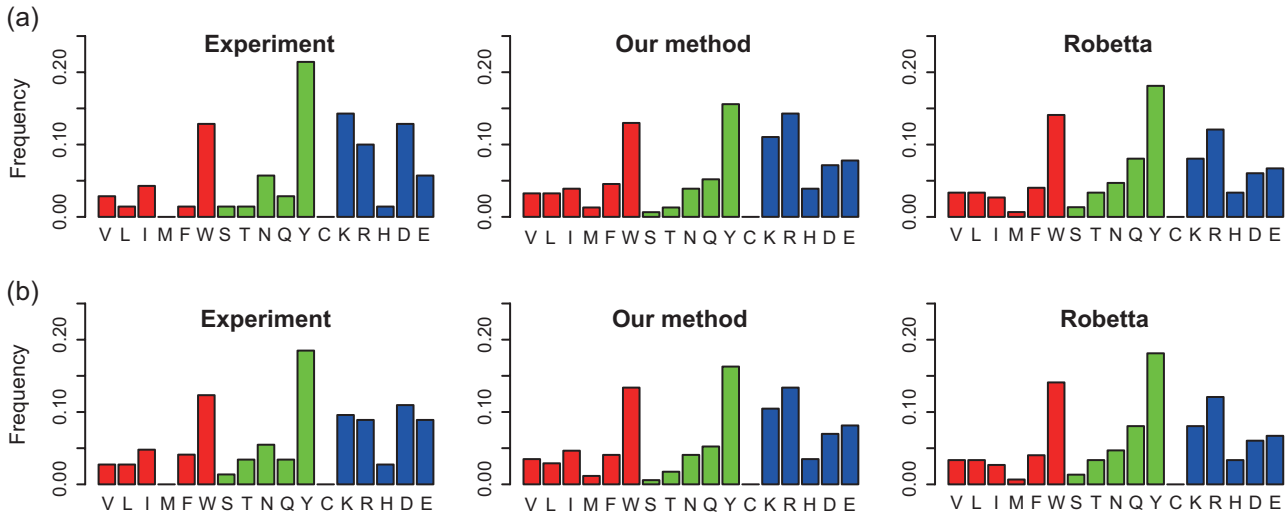

Fig. 5 Distribution representing how frequently each amino-acid species acts as a hot spot. The distribution is experimentally measured or theoretically predicted (our method or Robetta). Red, green, and blue colors represent nonpolar, polar, and charged residues, respectively. (a) Criterion 1 ($\theta_{\rm obs} = 2$ kcal/mol). (b) Criterion 2 ($\theta_{\rm obs} = 1$ kcal/mol).

Table 1 List of protein-protein complexes treated in the present study

PDB ID	Partner 1 (chain id)	Partner 2 (chain id)	Number of mutations	Number of hot spots from criterion 1	Number of hot spots from criterion 2
1a22	hGH (A)	hGHbp (B)	53	7	14
1a4y	RNase inhibitor (A)	Angiogenin (B)	24	3	6
1ahw	Fab 5G9 (A,B)	Tissue factor (C)	8	1	5
1brs	Barnase (A)	Barstar (D)	12	8	11
1bxi	Im9 (A)	E9 DNase (B)	16	6	8
1cbw	Cymotrypsin (A,B,C)	BPTI (D)	6	1	1
1dan	Factor VIIA (L,H)	Tissue factor (T,U)	64	2	10
1dfj	RNase A (E)	RNase inhibitor (I)	14	4	11
1dvf	FV D1.3 (A,B)	FV E5.2 (C,D)	25	9	22
1dx5	Thrombin (A,M)	Thrombomodulin (I)	16	5	8
1fcc	Protein G (A)	IGG (C)	8	4	5
1gc1	Envelope protein GP120 (G)	CD4 (C)	17	0	3
1jck	T-cell antigen receptor (A)	SEC3-1A4 (B)	9	4	8
1jrh	Antibody A6 (L,H)	Interferon γ receptor (I)	27	8	17
1nmb	N9 Neuraminidase (N)	Fab NC10 (L,H)	1	0	1
1vfb	D1.3 (A,B)	HEL (C)	26	3	10
2ptc	Trypsin (E)	BPTI (I)	1	1	1
3hfm	HYHEL-10 (L,H)	HEL (Y)	14	4	5

Table 3 Performance evaluated from criterion 1 ($\theta_{\rm obs} = 2$ kcal/mol)

	Best threshold	Recall	Precision	Specificity	F-measure
Our method	5.84	0.871	0.396	0.657	0.545
Robetta	0.98 kcal/mol	0.786	0.369	0.653	0.502

Table 4 Performance evaluated from criterion 2 ($\theta_{\rm obs} = 1$ kcal/mol)

	Best threshold	Recall	Precision	Specificity	F-measure
Our method	5.01	0.733	0.622	0.667	0.673
Robetta	0.98 kcal/mol	0.664	0.651	0.733	0.658